Allgemeines Lineares Modell

BSc Psychologie, SoSe 2024

Joram Soch

(13) Multiple Regression

Anwendungsszenario

Modellformulierung

Modellschätzung

Modellevaluation

Selbstkontrollfragen

Anwendungsszenario

Modellformulierung

Modellschätzung

Modellevaluation

Selbstkontrollfragen

Anwendungsszenario

- Generalisierung der einfachen linearen Regression zu mehr als einer unabhängigen Variable.
- Eine univariate abhängige Variable (AV) bestimmt an randomisierten experimentellen Einheiten.
- Zwei oder mehr "kontinuierliche" unabhängige Variablen (UV).
- Die unabhängigen Variablen heißen Regressoren, Prädiktoren oder Kovariaten.

Ziele

- Quantifizierung des Einflusses einzelner UVs auf die AV im Kontext anderer UVs.
- Quantifizierung des Erklärungspotentials der Variation der AV durch die Variation der UVs.
- Vorhersage von Werten der AV aus (neuen) Werten der UVs nach Parameterschätzung.

Anwendungsbeispiel

- BDI-Differenzwerte (`BDI`) in Abhängigkeit von Therapiedauer (`Duration`) und Alter (`Age`)

Anwendungsszenario

Beispieldatensatz ($n = 100$; hier: $i = 1, ..., 25$)

ID	Age	Duration	BDI
1	50	16	9
2	38	13	9
3	46	16	10
4	62	17	3
5	25	23	38
6	34	23	29
7	36	24	31
8	36	18	21
9	57	20	20
10	46	17	16
11	59	21	18
12	54	22	24
13	27	14	20
14	56	18	12
15	41	20	31
16	46	23	30
17	23	15	23
18	36	22	27
19	44	19	23
20	70	20	11
21	72	13	-3
22	57	16	12
23	67	16	3
24	41	19	23
25	44	14	15

Anwendungsszenario

Modellformulierung

Modellschätzung

Modellevaluation

Selbstkontrollfragen

Modellformulierung

Definition (Modell der multiplen Regression)

y_i mit $i = 1, ..., n$ sei die Zufallsvariable, die den $i$ten Wert einer abhängigen Variable modelliert. Dann hat das *Modell der multiplen Regression* die strukturelle Form

$$y_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, ..., n \text{ und } \sigma^2 > 0, \tag{1}$$

wobei $x_{ij} \in \mathbb{R}$ mit $1 \leq i \leq n$ und $1 \leq j \leq p$ den $i$ten Wert der $j$te unabhängigen Variable bezeichnet. Die unabhängigen Variablen werden auch *Regressoren*, *Prädiktoren* oder Kovariaten genannt. Mit

$$x_i := (x_{i1}, ..., x_{ip})^{\mathrm{T}} \in \mathbb{R}^p \text{ und } \beta := (\beta_1, ..., \beta_p)^{\mathrm{T}} \in \mathbb{R}^p \tag{2}$$

hat das Modell der multiplen Regression die Datenverteilungsform

$$y_i \sim N(\mu_i, \sigma^2) \text{ u.v. für } i = 1, ..., n, \text{ wobei } \mu_i := x_i^{\mathrm{T}}\beta. \tag{3}$$

In diesem Zusammenhang wird $x_i \in \mathbb{R}^p$ auch als $i$ter *Merkmalsvektor* bezeichnet. Die Designmatrixform des Modells der multiplen Regression schließlich ist gegeben durch

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{4}$$

mit

$$y := (y_1, ..., y_n)^{\mathrm{T}},\ X := (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p} \in \mathbb{R}^{n \times p},\ \beta := (\beta_1, ..., \beta_p)^{\mathrm{T}} \in \mathbb{R}^p \text{ und } \sigma^2 > 0. \tag{5}$$

Bemerkung

- Das Modell der multiplen Regression und die allgemeine Form des ALMs sind identisch.

Beispielmodell

Wir betrachen als Beispiel ein multiples Regressionsmodell mit drei Regressoren der Form

$$y = X\beta + \varepsilon, \varepsilon \sim N(0_n, \sigma^2 I_n) \text{ mit } X := \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \text{ und } \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}. \tag{6}$$

Zur Veranschaulichung können wir das Beispielmodell wie folgt ausformulieren.

$$\begin{aligned} & y = X\beta + \varepsilon \\ \Leftrightarrow & \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \\ \Leftrightarrow & \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + x_{11}\beta_1 + x_{12}\beta_2 + \varepsilon_1 \\ \beta_0 + x_{21}\beta_1 + x_{22}\beta_2 + \varepsilon_2 \\ \vdots \\ \beta_0 + x_{n1}\beta_1 + x_{n2}\beta_2 + \varepsilon_n \end{pmatrix} \end{aligned} \tag{7}$$

Mit diesem Modell erzeugen wir nachfolgend einen Beispieldatensatz mit 100 Datenpunkten.

Modellformulierung

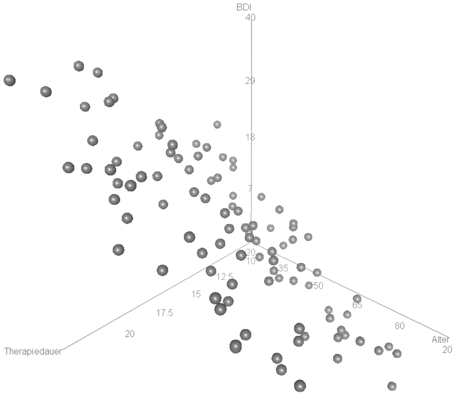
Beispieldatensatzerzeugung

```
# Datensimulation
library(MASS)                                          # multivariate Normalverteilung
set.seed(10)                                           # reproduzierbare Daten
n            = 100                                     # Anzahl Datenpunkte
p            = 3                                       # Anzahl Parameter
x_1          = round(runif(n,20,80))                   # Regressorwerte Alter
x_2          = round(runif(n,12,24))                   # Regressorwerte Therapiedauer
X            = matrix(c(rep(1,n),x_1,x_2), nrow = n)   # Designmatrix
I_n          = diag(n)                                 # Identitätsmatrix
beta         = matrix(c(5,-.5,2), nrow = p)            # Betaparametervektor
sigsqr       = 10                                      # Varianzparameter
y            = mvrnorm(1, X %*% beta, sigsqr*I_n)      # eine Realisierung des n-dimensionalen ZVs

# Dataframeformatierung
D            = data.frame("ID" = 1:n)                  # Dataframe-Initialisierung und ID-Variable
D$Age        = x_1                                     # Alter
D$Duration   = x_2                                     # Therapiedauer
D$BDI        = y                                       # Pre-Post-BDI-Differenzwerte

# Datenspeicherung
write.csv(D, file = file.path(getwd(), "Daten", "Multiple_Regression_Daten.csv"))
```

Beispieldatenvisualisierung

```r
# Daten einlesen
fname      = file.path(getwd(), "Daten", "Multiple_Regression_Daten.csv")
D          = read.table(fname, sep = ",", header = TRUE)

# Visualisierung mit der Funktion scatter3d() aus dem Package "car"
library(car)
scatter3d(D$Age, D$BDI, D$Duration,
  xlab        = "Alter",
  ylab        = "BDI",
  zlab        = "Therapiedauer",
  point.col   = "gray40",
  axis.col    = rep("black",3),
  axis.scales = T,
  axis.ticks  = T,
  surface     = F)
```

Beispieldatenvisualisierung



Anwendungsszenario

Modellformulierung

Modellschätzung

Modellevaluation

Selbstkontrollfragen

Modellschätzung

Überblick

Der Betaparameterschätzer hat bekanntlich die Form

$$\hat{\beta} := (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \tag{8}$$

Dabei quantifizieren in sehr grober Auflösung

- $X^{\mathrm{T}}y \in \mathbb{R}^{p}$ die Kovariabilität der Regressoren mit den Daten und
- $X^{\mathrm{T}}X \in \mathbb{R}^{p\times p}$ die Kovariabilität der Regressoren untereinander.

Damit ergibt sich für die Betaparameterschätzer also eine Interpretation als "regressorkovariabilitäts-normalisierte Regressordatenkovariabilität":

$$\hat{\beta} \approx \text{Regressorkovariabilität}^{-1} \cdot \text{Regresssordatenkovariabilität.} \tag{9}$$

Im Folgenden wollen wir diese Intuition am Beispiel einer einfachen multiplen Regression mit einem Interzeptregressor und zwei unabhängigen Variablen Regressoren vertiefen, wobei die betreffenden Kovariabilitäten einmal durch Stichprobenkorrelationen und einmal durch partielle Stichprobenkorrelationen quantifiziert werden sollen.

Theorem (Betaparameterschätzer und Korrelationen)

Gegeben sei ein multiples Regressionsmodel der Form

$$y = X\beta + \varepsilon, \varepsilon \sim N(0_n, \sigma^2 I_n) \text{ mit } X := \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \text{ und } \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}. \tag{10}$$

Dann gilt

$$\hat{\beta} = \begin{pmatrix} \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 \\ \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \frac{s_y}{s_{x_1}} \\ \frac{r_{y,x_2} - r_{y,x_1} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \frac{s_y}{s_{x_2}} \end{pmatrix}, \tag{11}$$

wobei $\bar{x}$ ein Stichprobenmittel, s_x eine Stichprobenstandardabweichunge und $r_{x,y}$ eine Stichprobenkorrelation für die entsprechenden y_i, x_{i1} und x_{i2} mit $i = 1, ..., n$ bezeichnet.

Bemerkungen

- In Bezug auf die Regressoren sind die Begriffe Stichprobenmittel, Stichprobenstandardabweichung, und Stichprobenkorrelation lediglich formal gemeint, nach Voraussetzung des ALMs sind die Regressorenwerte keine Realisierungen von Zufallsvariablen.

Modellschätzung

Bemerkungen (fortgeführt)

Exemplarisch betrachten wir

$$\hat{\beta}_1 = \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \frac{s_y}{s_{x_1}}. \tag{12}$$

Man erkennt unter anderem:

- Im Fall $r_{x_1,x_2} = 0$ und $s_y = s_{x_1}$ gilt $\hat{\beta}_1 = r_{y,x_1}$.
- Im Fall maximaler Korrelation $r_{x_1,x_2} = \pm 1$ ist $\hat{\beta}_1$ nicht definiert.
- Je größer $|r_{x_1,x_2}|$, desto größer der von r_{y,x_1} subtrahierte Term $r_{y,x_2} r_{x_1,x_2}$.
- Je größer $|r_{y,x_2}|$, desto größer der von r_{y,x_1} subtrahierte Term $r_{y,x_2} r_{x_1,x_2}$.
- Je größer also die Korrelation von x_2 mit x_1 oder die Korrelation von x_2 mit den Daten ist, desto mehr wird von dem Term r_{y,x_1} abgezogen, der die Kovariabilität von x_1 mit den Daten repräsentiert.
- Bei identischen Korrelationen und gleich bleibender Regressorstandabweichung steigt $\hat{\beta}_1$ mit s_y.

<u>Beweis</u>

Wir erinnern zunächst daran, dass die Form des Betaparameterschätzers bekanntlich zum System der Normalengleichungen äquivalent ist (vgl. Einheit (6) Parameterschätzung):

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \Leftrightarrow X^{\mathrm{T}}X\hat{\beta} = X^{\mathrm{T}}y. \tag{13}$$

Ausschreiben des Normalengleichungssystems für den hier betrachteten Spezialfall des ALMs ergibt dann zunächst

$$
\begin{aligned}
& X^{\mathrm{T}}X\hat{\beta} = X^{\mathrm{T}}y \\
\Leftrightarrow & \begin{pmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\
\Leftrightarrow & \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_{i1} \\ \sum_{i=1}^n y_i x_{i2} \end{pmatrix}
\end{aligned}
$$

Modellschätzung

Beweis (fortgeführt)

und damit

$$
\begin{aligned}
& X^{\mathrm{T}} X \hat{\beta} = X^{\mathrm{T}} y \\
\Leftrightarrow & \begin{pmatrix} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} \\ \hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1}x_{i2} \\ \hat{\beta}_0 \sum_{i=1}^n x_{i2} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}x_{i2} + \hat{\beta}_2 \sum_{i=1}^n x_{i2}^2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_{i1} \\ \sum_{i=1}^n y_i x_{i2} \end{pmatrix}
\end{aligned}
$$

Aus der Gleichung der ersten Vektorkomponenten folgt dann direkt die Form von $\hat{\beta}_0$ mit

$$
\begin{aligned}
& \sum_{i=1}^n y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} \\
\Leftrightarrow\ & \frac{1}{n}\sum_{i=1}^n y_i = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^n x_{i1} + \hat{\beta}_2 \frac{1}{n}\sum_{i=1}^n x_{i2} \\
\Leftrightarrow\ & \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2
\end{aligned} \tag{14}
$$

Beweis (fortgeführt)

Einsetzen dieser Form von $\hat{\beta}_0$ in die Gleichung der zweiten Vektorkomponenten ergibt dann

$$
\begin{aligned}
\hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1} x_{i2} &= \sum_{i=1}^n y_i x_{i1} \\
\Leftrightarrow (\bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2) \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1} x_{i2} &= \sum_{i=1}^n y_i x_{i1} \\
\Leftrightarrow \bar{y} \sum_{i=1}^n x_{i1} - \hat{\beta}_1 \bar{x}_1 \sum_{i=1}^n x_{i1} - \hat{\beta}_2 \bar{x}_2 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1} x_{i2} &= \sum_{i=1}^n y_i x_{i1} \\
\Leftrightarrow \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 - \hat{\beta}_1 \bar{x}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i1} x_{i2} - \hat{\beta}_2 \bar{x}_2 \sum_{i=1}^n x_{i1} &= \sum_{i=1}^n y_i x_{i1} - \bar{y} \sum_{i=1}^n x_{i1} \\
\Leftrightarrow \hat{\beta}_1 \left(\sum_{i=1}^n x_{i1}^2 - \bar{x}_1 \sum_{i=1}^n x_{i1} \right) + \hat{\beta}_2 \left(\sum_{i=1}^n x_{i1} x_{i2} - \bar{x}_2 \sum_{i=1}^n x_{i1} \right) &= \sum_{i=1}^n y_i x_{i1} - \bar{y} \sum_{i=1}^n x_{i1}
\end{aligned}
$$

Beweis (fortgeführt)

Im Beweis des Theorems zur Ausgleichsgerade (vgl. Einheit (1) Regression) haben wir gesehen, dass

$$\sum_{i=1}^{n} x_{i1}x_{i1} - \bar{x}_1 \sum_{i=1}^{n} x_{i1} = \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i1} - \bar{x}_1), \tag{15}$$

$$\sum_{i=1}^{n} x_{i1}x_{i2} - \bar{x}_2 \sum_{i=1}^{n} x_{i1} = \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \text{ und} \tag{16}$$

$$\sum_{i=1}^{n} y_i x_{i1} - \bar{y} \sum_{i=1}^{n} x_{i1} = \sum_{i=1}^{n}(y_i - \bar{y})(x_{i1} - \bar{x}_1) \tag{17}$$

Modellschätzung

Beweis (fortgeführt)

Es ergibt sich also, dass

$$
\begin{aligned}
\hat{\beta}_1 \sum_{i=1}^n (x_{i1}-\bar{x}_1)(x_{i1}-\bar{x}_1) + \hat{\beta}_2 \sum_{i=1}^n (x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2) &= \sum_{i=1}^n (y_i-\bar{y})(x_{i1}-\bar{x}_1) \\
\Leftrightarrow \hat{\beta}_1 \frac{\sum_{i=1}^n (x_{i1}-\bar{x}_1)(x_{i1}-\bar{x}_1)}{n-1} + \hat{\beta}_2 \frac{\sum_{i=1}^n (x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{n-1} &= \frac{\sum_{i=1}^n (y_i-\bar{y})(x_{i1}-\bar{x}_1)}{n-1}.
\end{aligned}
\tag{18}
$$

Mit den Definitionen von Stichprobenstandardabweichung und -korrelation folgt dann weiter

$$
\begin{aligned}
\hat{\beta}_1 s_{x_1} s_{x_1} + \hat{\beta}_2 c_{x_1,x_2} &= c_{y,x_1} \\
\Leftrightarrow \hat{\beta}_1 \frac{s_{x_1} s_{x_1}}{s_y s_{x_1}} + \hat{\beta}_2 \frac{c_{x_1,x_2}}{s_y s_{x_1}} &= \frac{c_{y,x_1}}{s_y s_{x_1}} \\
\Leftrightarrow \hat{\beta}_1 \frac{s_{x_1}}{s_y} + \hat{\beta}_2 \frac{c_{x_1,x_2}}{s_y s_{x_1}} &= r_{y,x_1} \\
\Leftrightarrow \hat{\beta}_1 \frac{s_{x_1}}{s_y} + \hat{\beta}_2 \frac{c_{x_1,x_2} s_{x_2}}{s_y s_{x_1} s_{x_2}} &= r_{y,x_1} \\
\Leftrightarrow \hat{\beta}_1 \frac{s_{x_1}}{s_y} + \hat{\beta}_2 \frac{s_{x_2}}{s_y} r_{x_1,x_2} &= r_{y,x_1}
\end{aligned}
\tag{19}
$$

Beweis (fortgeführt)

Definition von

$$b_j := \hat{\beta}_j \frac{s_{x_j}}{s_y},\ j = 1,2 \tag{20}$$

erlaubt dann die Schreibweise

$$b_1 + b_2 r_{x_1,x_2} = r_{y,x_1}. \tag{21}$$

Durch Vertauschen der Subskripte folgt analog aus der Gleichung der dritten Vektorkomponenten

$$b_1 r_{x_1,x_2} + b_2 = r_{y,x_2}. \tag{22}$$

Insgesamt haben wir also gesehen, dass die Definition des Betaparameterschätzers im vorliegenden ALM-Spezialfall ergibt, dass mit

$$\hat{\beta}_j = b_j \frac{s_y}{s_{x_j}},\ j = 1,2 \tag{23}$$

gilt, dass

$$\begin{aligned} r_{y,x_1} &= b_1 + b_2 r_{x_1,x_2} \\ r_{y,x_2} &= b_1 r_{x_1,x_2} + b_2 \end{aligned} \tag{24}$$

Modellschätzung

Beweis (fortgeführt)

Damit folgt aus der zweiten Gleichung dann sofort

$$b_2 = r_{y,x_2} - b_1 r_{x_1,x_2}. \tag{25}$$

Einsetzen in die erste Gleichung ergibt dann

$$\begin{aligned}
b_1 + (r_{y,x_2} - b_1 r_{x_1,x_2}) r_{x_1,x_2} &= r_{y,x_1} \\
\Leftrightarrow b_1 + r_{y,x_2} r_{x_1,x_2} - b_1 r_{x_1,x_2}^2 &= r_{y,x_1} \\
\Leftrightarrow r_{y,x_2} r_{x_1,x_2} + b_1 \left(1 - r_{x_1,x_2}^2\right) &= r_{y,x_1} \\
\Leftrightarrow b_1 \left(1 - r_{x_1,x_2}^2\right) &= r_{y,x_1} - r_{y,x_2} r_{x_1,x_2} \\
\Leftrightarrow b_1 &= \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2}
\end{aligned} \tag{26}$$

Beweis (fortgeführt)

Für b_2 ergibt sich damit weiterhin

$$
\begin{aligned}
b_2 &= r_{y,x_2} - b_1 r_{x_1,x_2} \\
\Leftrightarrow b_2 &= r_{y,x_2} - \left(\frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}}\right) r_{x_1,x_2} \\
\Leftrightarrow b_2 &= \frac{r_{y,x_2}\left(1 - r^2_{x_1,x_2}\right)}{1 - r^2_{x_1,x_2}} - \frac{r_{y,x_1} r_{x_1,x_2} - r_{y,x_2} r^2_{x_1,x_2}}{1 - r^2_{x_1,x_2}} \\
\Leftrightarrow b_2 &= \frac{r_{y,x_2} - r_{y,x_2} r^2_{x_1,x_2} - r_{y,x_1} r_{x_1,x_2} + r_{y,x_2} r^2_{x_1,x_2}}{1 - r^2_{x_1,x_2}} \\
\Leftrightarrow b_2 &= \frac{r_{y,x_2} - r_{y,x_1} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}}.
\end{aligned}
\tag{27}
$$

Damit folgen dann aber

$$
\begin{aligned}
\hat{\beta}_1 &= b_1 \frac{s_y}{s_{x_1}} = \left(\frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}}\right) \frac{s_y}{s_{x_1}} \\
\hat{\beta}_2 &= b_2 \frac{s_y}{s_{x_2}} = \left(\frac{r_{y,x_2} - r_{y,x_1} r_{x_1,x_2}}{1 - r^2_{x_1,x_2}}\right) \frac{s_y}{s_{x_2}}
\end{aligned}
\tag{28}
$$

und es ist alles gezeigt. □

Anwendungsbeispiel

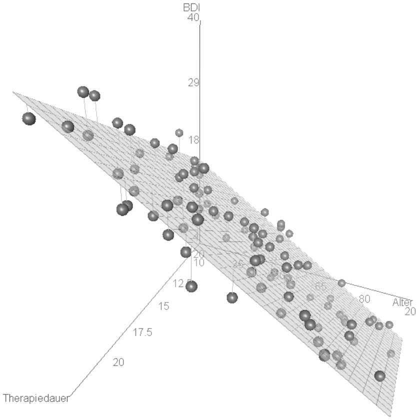
```
# Daten einlesen
fname      = file.path(getwd(), "Daten", "Multiple_Regression_Daten.csv")
D          = read.table(fname, sep = ",", header = TRUE)

# Modellschätzung
y          = D$BDI                                                 # abhängige Variable
n          = length(y)                                             # Anzahl Datenpunkte
X          = matrix(c(rep(1,n), D$Age, D$Duration), nrow = n)      # Designmatrix
beta_hat   = solve(t(X) %*% X) %*% t(X) %*% y                      # Betaparameterschätzer
eps_hat    = y - X %*% beta_hat                                    # Residuenvektor
sigsqr_hat = (t(eps_hat) %*% eps_hat) /(n-p)                       # Varianzparameterschätzer

# Betaparameterschätzer aus Stichprobenmittel, -standardabweichungen und -korrelationen
y12        = cbind(y,X[,-1])                                       # Matrix (y, x_1, x_2)
bars       = apply(y12, 2, mean)                                   # Stichprobenmittel
s          = apply(y12, 2, sd)                                     # Stichprobenstandardabweichungen
r          = cor(y12)                                              # Stichprobenkorrelationen
beta_hat_1 = (r[1,2] - r[1,3]*r[2,3])/(1 - r[2,3]^2)*(s[1]/s[2]) # \hat{\beta}_1
beta_hat_2 = (r[1,3] - r[1,2]*r[2,3])/(1 - r[2,3]^2)*(s[1]/s[3]) # \hat{\beta}_2
beta_hat_0 = bars[1] - beta_hat_1*bars[2] - beta_hat_2*bars[3]   # \hat{\beta}_0

# Ausgabe
cat(  "beta_hat ALM-Schätzer           :", beta_hat,
    "\nbeta_hat Deskriptivstatistiken :", c(beta_hat_0,beta_hat_1,beta_hat_2))
```

```
> beta_hat ALM-Schätzer           : 5.42 -0.481 1.91
> beta_hat Deskriptivstatistiken : 5.42 -0.481 1.91
```

Beispieldatenvisualisierung

```r
# Daten einlesen
fname      = file.path(getwd(), "Daten", "Multiple_Regression_Daten.csv")
D          = read.table(fname, sep = ",", header = TRUE)

# Visualisierung mit der Funktion scatter3d() aus dem Package "car"
library(car)
scatter3d(D$Age, D$BDI, D$Duration,
    xlab        = "Alter",
    ylab        = "BDI",
    zlab        = "Therapiedauer",
    point.col   = "gray40",
    axis.col    = rep("black",3),
    axis.scales = T,
    axis.ticks  = T,
    surface     = T,
    surface.col = "gray70",
    neg.res.col = "gray70",
    pos.res.col = "gray70")
```

Beispieldatenvisualisierung



Theorem (Betaparameterschätzer und partielle Korrelationen)

Gegeben sei ein multiples Regressionsmodel der Form

$$y = X\beta + \varepsilon, \varepsilon \sim N(0_n, \sigma^2 I_n) \text{ mit } X := \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \text{ und } \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}. \quad (29)$$

Dann gilt

$$\hat{\beta} = \begin{pmatrix} \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 \\ r_{y,x_1 \backslash x_2} \sqrt{\frac{1-r^2_{y,x_2}}{1-r^2_{x_1,x_2}}} \frac{s_y}{s_{x_1}} \\ r_{y,x_2 \backslash x_1} \sqrt{\frac{1-r^2_{y,x_1}}{1-r^2_{x_2,x_1}}} \frac{s_y}{s_{x_2}} \end{pmatrix}, \quad (30)$$

wobei für $1 \leq k, l \leq 2$ und $i = 1, ..., n$

- $r_{y,x_k \backslash x_l}$ die partielle Stichprobenkorrelation der y_i und x_{ik}, gegeben die x_{il} ist,
- r_{y,x_k} die Stichprobenkorrelation der y_i und x_{ik} ist, und
- r_{x_k,x_l} die Stichprobenkorrelation der x_{ik} und x_{il} ist.

Modellschätzung

Bemerkungen

- Wir betrachten exemplarisch

$$\hat{\beta}_1 = r_{y,x_1\backslash x_2}\sqrt{\frac{1-r_{y,x_2}^2}{1-r_{x_1,x_2}^2}}\frac{s_y}{s_{x_1}}. \tag{31}$$

- Im Allgemeinen gilt, dass $\hat{\beta}_1 \neq r_{y,x_1\backslash x_2}$.
- Betaparameterschätzer sind also im Allgemeinen keine partiellen Stichprobenkorrelationen.
- $\hat{\beta}_1 = r_{y,x_1\backslash x_2}$ gilt genau dann, wenn $s_y = s_{x_1}$ und zudem
    - $r_{y,x_2} = r_{x_1,x_2} = 0$, wenn also die Stichprobenkorrelationen der Daten und der Werte des zweiten Regressors sowie die Stichprobenkorrelation der Werte der beiden Regressoren gleich Null sind. Dies kann der Fall sein, wenn einer der Regressoren die Daten "sehr gut erklärt" und der andere Regressor von dem ersten "sehr verschieden" ist.
    - $|r_{y,x_2}| = |r_{x_1,x_2}|$, wenn also die obigen Stichprobenkorrelationen dem Betrage nach gleich sind. Dies ist vermutlich selten der Fall.

Beweis

Wir betrachten $\hat{\beta}_1$, das Resultat für $\hat{\beta}_2$ folgt dann durch Vertauschen der Indizes. Wir haben in vorherigem Theorem gesehen, dass

$$\hat{\beta}_1 = \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \frac{s_y}{s_{x_1}}. \tag{32}$$

Weiterhin haben wir bereits gesehen (vgl. Einheit (12) Partielle Korrelation), dass unter der Annahme der multivariaten Normalverteilung von y, x_1, x_2 ein Schätzer für die partielle Korrelation von y und x_1, gegeben x_2, durch

$$r_{y,x_1 \setminus x_2} = \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{\sqrt{1 - r_{y,x_2}^2}\sqrt{1 - r_{x_1,x_2}^2}} \tag{33}$$

gegeben ist. Für $\hat{\beta}_1$ ergibt sich somit

$$\begin{aligned}
\hat{\beta}_1 &= \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \frac{s_y}{s_{x_1}} \\
\Leftrightarrow \left(1 - r_{x_1,x_2}^2\right) \hat{\beta}_1 &= \left(r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}\right) \frac{s_y}{s_{x_1}} \\
\Leftrightarrow \frac{1 - r_{x_1,x_2}^2}{\sqrt{1 - r_{y,x_2}^2}\sqrt{1 - r_{x_1,x_2}^2}} \hat{\beta}_1 &= \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{\sqrt{1 - r_{y,x_2}^2}\sqrt{1 - r_{x_1,x_2}^2}} \frac{s_y}{s_{x_1}} \\
\Leftrightarrow \frac{1 - r_{x_1,x_2}^2}{\sqrt{1 - r_{y,x_2}^2}\sqrt{1 - r_{x_1,x_2}^2}} \hat{\beta}_1 &= r_{y,x_1 \setminus x_2} \frac{s_y}{s_{x_1}}
\end{aligned} \tag{34}$$

Beweis

und damit weiter

$$
\begin{aligned}
\hat{\beta}_1 &= r_{y,x_1\backslash x_2} \frac{\sqrt{1-r_{y,x_2}^2}\sqrt{1-r_{x_1,x_2}^2}}{1-r_{x_1,x_2}^2} \frac{s_y}{s_{x_1}} \\
\Leftrightarrow \hat{\beta}_1 &= r_{y,x_1\backslash x_2} \frac{\sqrt{1-r_{y,x_2}^2}\sqrt{1-r_{x_1,x_2}^2}}{\left(\sqrt{1-r_{x_1,x_2}^2}\right)^2} \frac{s_y}{s_{x_1}} \\
\Leftrightarrow \hat{\beta}_1 &= r_{y,x_1\backslash x_2} \frac{\sqrt{1-r_{y,x_2}^2}}{\sqrt{1-r_{x_1,x_2}^2}} \frac{s_y}{s_{x_1}} \\
\Leftrightarrow \hat{\beta}_1 &= r_{y,x_1\backslash x_2} \sqrt{\frac{1-r_{y,x_2}^2}{1-r_{x_1,x_2}^2}} \frac{s_y}{s_{x_1}}.
\end{aligned}
\tag{35}
$$

□

Modellschätzung

Anwendungsbeispiel

```
# Daten einlesen
fname       = file.path(getwd(), "Daten", "Multiple_Regression_Daten.csv")
D           = read.table(fname, sep = ",", header = TRUE)

# Modellschätzung
y           = D$BDI                                             # abhängige Variable
n           = length(y)                                         # Anzahl Datenpunkte
X           = matrix(c(rep(1,n), D$Age, D$Duration), nrow = n)  # Designmatrix
p           = ncol(X)                                           # Anzahl Parameter
beta_hat    = solve(t(X) %*% X) %*% t(X) %*% y                  # Betaparameterschätzer
eps_hat     = y - X %*% beta_hat                                # Residuenvektor
sigsqr_hat  = (t(eps_hat) %*% eps_hat) /(n-p)                   # Varianzparameterschätzer

# Betaparameterschätzer aus (partiellen) Korrelationen
library(ppcor)                                                  # Tools für partielle Korrelationen
y12         = cbind(y,X[,-1])                                   # Matrix (y, x_1, x_2)
bars        = apply(y12, 2, mean)                               # Stichprobenmittel
s           = apply(y12, 2, sd)                                 # Stichprobenstandardabweichungen
r           = cor(y12)                                          # Stichprobenkorrelationen
pr          = pcor(y12)                                         # partielle Stichprobenkorrelationen
pr          = pr$estimate                                       # partielle Stichprobenkorrelationen
beta_hat_1 = pr[1,2]*sqrt((1-r[1,3]^2)/(1-r[2,3]^2))*(s[1]/s[2]) # \hat{\beta}_1
beta_hat_2 = pr[1,3]*sqrt((1-r[1,2]^2)/(1-r[3,2]^2))*(s[1]/s[3]) # \hat{\beta}_2
beta_hat_0 = bars[1] - beta_hat_1*bars[2] - beta_hat_2*bars[3]   # \hat{\beta}_0

# Ausgabe
cat(  "Korrelationen r(y,x_1), r(y,x_2), r(x_1,x_2)       :", c(r[1,2],r[1,3],r[2,3]),
    "\npartielle Korrelationen r(y,x_1|x_2), r(y,x_2|x_1) :", c(pr[1,2],pr[1,3]),
    "\nbeta_hat ALM-Schätzer                              :", beta_hat,
    "\nbeta_hat partielle Korrelationen                   :", c(beta_hat_0,beta_hat_1,beta_hat_2))
```

```
> Korrelationen r(y,x_1), r(y,x_2), r(x_1,x_2)       : -0.726 0.644 -0.0268
> partielle Korrelationen r(y,x_1|x_2), r(y,x_2|x_1) : -0.927 0.909
> beta_hat ALM-Schätzer                              : 5.42 -0.481 1.91
> beta_hat partielle Korrelationen                   : 5.42 -0.481 1.91
```

Anwendungsszenario

Modellformulierung

Modellschätzung

Modellevaluation

Selbstkontrollfragen

Parameterinferenz: T-Tests

Theorem (Verteilung der T-Statistik)

Gegeben seien das ALM

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \tag{36}$$

sowie die Betaparameter- und Varianzparameterschätzer

$$\hat{\beta} := (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \text{ und } \hat{\sigma}^2 := \frac{(y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})}{n-p}\ . \tag{37}$$

Schließlich sei für einen Kontrastgewichtsvektor $c \in \mathbb{R}^p$ und einen Nullparameter $\beta_0 \in \mathbb{R}^p$ die T-Statistik

$$T := \frac{c^{\mathrm{T}}\hat{\beta} - c^{\mathrm{T}}\beta_0}{\sqrt{\hat{\sigma}^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}}\ . \tag{38}$$

Dann gilt:

$$T \sim t(\delta, n-p) \text{ mit } \delta = \frac{c^{\mathrm{T}}\beta - c^{\mathrm{T}}\beta_0}{\sqrt{\sigma^2 c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c}}. \tag{39}$$

Bemerkungen

- Das Theorem wurde bereits eingeführt (siehe Einheit (7) in *Allgemeines Lineares Modell*).

Parameterinferenz: T-Tests

Einige mögliche Kontrastgewichtsvektoren und Nullhypothesen im Anwendungsbeispiel sind:

$$
\begin{array}{lll}
c = (1,0,0)^{\mathrm{T}} & H_0 : \beta_0 = 0 & H_A : \beta_0 \neq 0 \\
c = (0,1,0)^{\mathrm{T}} & H_0 : \beta_1 = 0 & H_A : \beta_1 \neq 0 \\
c = (0,0,1)^{\mathrm{T}} & H_0 : \beta_2 = 0 & H_A : \beta_2 \neq 0 \\
c = (0,1,-1)^{\mathrm{T}} & H_0 : \beta_1 - \beta_2 = 0 & H_A : \beta_1 - \beta_2 \neq 0 \\
c = (0,-1,1)^{\mathrm{T}} & H_0 : \beta_2 - \beta_1 = 0 & H_A : \beta_2 - \beta_1 \neq 0 \\
\ldots & \ldots & \ldots
\end{array}
$$

Parameterinferenz: T-Tests

```
# Daten einlesen
fname      = file.path(getwd(), "Daten", "Multiple_Regression_Daten.csv")
D          = read.table(fname, sep = ",", header = TRUE)

# Modellschätzung
y          = D$BDI                                            # abhängige Variable
n          = length(y)                                        # Anzahl Datenpunkte
X          = matrix(c(rep(1,n), D$Age, D$Duration), nrow = n) # Designmatrix
p          = ncol(X)                                          # Anzahl Parameter
beta_hat   = solve(t(X) %*% X) %*% t(X) %*% y                 # Betaparameterschätzer
eps_hat    = y - X %*% beta_hat                               # Residuenvektor
sigsqr_hat = (t(eps_hat) %*% eps_hat)/(n-p)                   # Varianzparameterschätzer

# Modellevaluation und Parameterinferenz
C          = cbind(diag(p), matrix(c(0,1,-1), nrow = 3))      # Kontrastgewichtsvektoren
ste        = rep(NaN, ncol(C))                                # Konstraststandardfehler
tee        = rep(NaN, ncol(C))                                # T-Statistiken
pvals      = rep(NaN, ncol(C))                                # p-Werte
for(i in 1:ncol(C)){
    c        = C[,i]                                          # Kontrastgewichtsvektor
    t_num    = t(c)%*%beta_hat                                # Zähler der T-Statistik
    ste[i]   = sqrt(sigsqr_hat*t(c)%*%solve(t(X)%*%X)%*%c)    # Kontraststandardfehler/Nenner der T-Statistik
    tee[i]   = t_num/ste[i]                                   # T-Statistik
    pvals[i] = 2*(1 - pt(abs(tee[i]),n-p))                    # p-Wert
}
# Ausgabe
R           = data.frame(c(beta_hat, t(C[,4]%*%beta_hat)), ste, tee, pvals)
rownames(R)  = c("(Intercept)", "Age", "Therapy", "Age-Therapy")
colnames(R)  = c("Estimate", "Std. Error", "t-value", "Pr(>|t|)")
print(R)
```

```
>              Estimate Std. Error t-value Pr(>|t|)
> (Intercept)     5.422     1.9024    2.85  0.00534
> Age            -0.481     0.0198  -24.33  0.00000
> Therapy         1.912     0.0893   21.41  0.00000
> Age-Therapy    -2.393     0.0909  -26.32  0.00000
```

Modellinferenz: F-Tests

Theorem (Verteilung der F-Statistik)

Gegeben seien das ALM

$$y = X\beta + \varepsilon \text{ mit } \varepsilon \sim N(0_n, \sigma^2 I_n) \,, \tag{40}$$

wobei $X \in \mathbb{R}^{n \times p}$, und mit $p = p_0 + p_1$ eine Partitionierung

$$\begin{aligned} X &= \begin{pmatrix} X_0 & X_1 \end{pmatrix},\; X_0 \in \mathbb{R}^{n \times p_0},\; X_1 \in \mathbb{R}^{n \times p_1} \text{ und} \\ \beta &= \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},\; \beta_0 \in \mathbb{R}^{p_0},\; \beta_1 \in \mathbb{R}^{p_1} \,. \end{aligned} \tag{41}$$

Schließlich sei ein Vektor c gegeben durch

$$c := \begin{pmatrix} 0_{p_0} \\ 1_{p_1} \end{pmatrix} \in \mathbb{R}^p \,. \tag{42}$$

Dann gilt

$$F \sim f(\delta, p_1, n-p) \text{ mit } \delta := \frac{c^{\mathrm{T}}\beta \left(c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}c\right)^{-1} c^{\mathrm{T}}\beta}{\sigma^2} \tag{43}$$

Bemerkungen

- Das Theorem wurde bereits eingeführt (siehe Einheit (8) in *Allgemeines Lineares Modell*).

Modellinferenz: F-Tests

Einige mögliche Partitionierungen und Nullhypothesen im Anwendungsbeispiel sind:

$$
\begin{array}{llll}
X_0 = \begin{pmatrix} 1_n \end{pmatrix} & X_1 = \begin{pmatrix} x_1 & x_2 \end{pmatrix} & H_0 : \beta_1 = 0 \wedge \beta_2 = 0 & H_A : \beta_1 \neq 0 \vee \beta_2 \neq 0 \\
X_0 = \begin{pmatrix} 1_n & x_1 \end{pmatrix} & X_1 = \begin{pmatrix} x_2 \end{pmatrix} & H_0 : \beta_2 = 0 & H_A : \beta_2 \neq 0 \\
X_0 = \begin{pmatrix} 1_n & x_2 \end{pmatrix} & X_1 = \begin{pmatrix} x_1 \end{pmatrix} & H_0 : \beta_1 = 0 & H_A : \beta_1 \neq 0 \\
X_0 = \begin{pmatrix} x_1 & x_2 \end{pmatrix} & X_1 = \begin{pmatrix} 1_n \end{pmatrix} & H_0 : \beta_0 = 0 & H_A : \beta_0 \neq 0
\end{array}
$$

Modellevaluation

Modellinferenz: F-Tests

```
# Daten einlesen
fname       = file.path(getwd(), "Daten", "Multiple_Regression_Daten.csv")
D           = read.table(fname, sep = ",", header = TRUE)

# Modellevaluation
y           = D$BDI                                                # abhängige Variable
n           = length(y)                                            # Anzahl Datenpunkte
X           = matrix(c(rep(1,n), D$Age, D$Duration), nrow = n) # Desigmatrix vollständiges Modell
p           = ncol(X)                                              # Anzahl Parameter vollständiges Modell
p_0         = 1        # reduziertes Modell:   (1_n)               # Anzahl Parameter reduziertes Modell
p_1         = p - p_0  # vollständiges Modell: (1_n, x_1, x_2) # Anzahl zusätzlicher Parameter im vollst. Modell
X_0         = X[,1:p_0]                                            # Designmatrix reduzierters Modell
beta_hat_0 = solve(t(X_0)%*%X_0)%*%t(X_0)%*%y                      # Betaparameterschätzer reduziertes Modell
beta_hat    = solve(t(X) %*%X )%*%t(X) %*%y                        # Betaparameterschätzer vollständiges Modell
eps_hat_0   = y - X_0 %*% beta_hat_0                               # Residuenvektor reduziertes Modell
eps_hat     = y - X %*% beta_hat                                   # Residuenvektor vollständiges Modell
eh0_eh0     = t(eps_hat_0) %*% eps_hat_0                           # residuelle Quadratsumme reduziertes Modell
eh_eh       = t(eps_hat) %*% eps_hat                               # residuelle Quadratsumme vollständiges Modell
sigsqr_hat = eh_eh/(n-p)                                           # Varianzparameterschätzer vollständiges Modell
f           = ((eh0_eh0-eh_eh)/p_1)/sigsqr_hat                     # F-Statistik
pval        = 1 - pf(f,p_1,n-p)                                    # p-Wert

# Ausgabe
cat("F-statistic: ", f, "on", p_1, "and", n-p, "DF,", " p-value: ", paste(pval))
```

```
> F-statistic:  540 on 2 and 97 DF,  p-value:  0
```

Modellformulierung, Modellschätzung und Modellevaluation mit R

```r
fname  = file.path(getwd(), "Daten", "Multiple_Regression_Daten.csv") # Dateiname
D      = read.table(fname, sep = ",", header = TRUE)                  # Datenframe
alm    = lm(BDI ~ Age + Duration, data = D)                           # Modellformulierung und Modellschätzung
summary(alm)
```

```
>
> Call:
> lm(formula = BDI ~ Age + Duration, data = D)
>
> Residuals:
>    Min     1Q Median     3Q    Max
> -7.178 -2.165  0.438  2.585  7.119
>
> Coefficients:
>             Estimate Std. Error t value Pr(>|t|)
> (Intercept)   5.4225     1.9024    2.85   0.0053 **
> Age          -0.4815     0.0198  -24.33   <2e-16 ***
> Duration      1.9119     0.0893   21.41   <2e-16 ***
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 3.07 on 97 degrees of freedom
> Multiple R-squared:  0.918,   Adjusted R-squared:  0.916
> F-statistic:  540 on 2 and 97 DF,  p-value: <2e-16
```

Anwendungsszenario

Modellformulierung

Modellschätzung

Modellevaluation

Selbstkontrollfragen

Selbstkontrollfragen

1. Erläutern Sie das Anwendungsszenario und die Ziele der multiplen Regression.
2. Definieren Sie das Modell der multiplen Regression.
3. Erläutern Sie die Begriffe Regressor, Prädiktor und Kovariate im Rahmen der multiplen Regression.
4. Erläutern Sie, warum $\hat{\beta} \approx \text{Regressorkovariabilität}^{-1} \cdot \text{Regressordatenkovariabilität}$ gilt.
5. Erläutern Sie den Zusammenhang zwischen Betaparameterschätzern und Korrelationen in einem multiplen Regressionmodell mit Interzeptparameter und zwei kontinuierlichen Prädiktoren anhand der Formel

$$\hat{\beta}_1 = \frac{r_{y,x_1} - r_{y,x_2} r_{x_1,x_2}}{1 - r_{x_1,x_2}^2} \frac{s_y}{s_{x_1}}. \tag{44}$$

6. Erläutern Sie den Zusammenhang zwischen Betaparameterschätzern und partieller Korrelation in einem multiplen Regressionmodell mit Interzeptparameter und zwei kontinuierlichen Prädiktoren anhand der Formel

$$\hat{\beta}_1 = r_{y,x_1 \backslash x_2} \sqrt{\frac{1 - r_y^2, x_2}{1 - r_{x_1,x_2}^2}} \frac{s_y}{s_{x_1}}. \tag{45}$$

7. $X \in \mathbb{R}^{n \times 2}$ sei die Designmatrix eines multiplen Regressionsmodells mit zwei Prädiktoren und Betaparametervektor $\beta := (\beta_1, \beta_2)^{\mathrm{T}}$. Geben Sie den Kontrastgewichtsvektor an, um die Nullhypothese $H_0 : \beta_1 = \beta_2$ mithilfe der T-Statistik zu testen.